Prediction of Structured Non-Coding RNAs in the Genomes of the Nematodes *Caenorhabditis elegans* and *Caenorhabditis briggsae*

Kristin Missal* [a], Xiaopeng Zhu [b], Dominic Rose [a], Wei Deng* [b], Geir Skogerbø [b], Runsheng Chen [b,c,d], and Peter F. Stadler [a,e,f]

[a] *Bioinformatics Group, Department of Computer Science, and Interdisciplinary Center for Bioinformatics, University of Leipzig, Härtelstraße 16-18, D-04107 Leipzig, Germany*
{kristin,studla}@bioinf.uni-leipzig.de

[b] *Bioinformatics Laboratory, Institute of Biophysics, Chinese Academy of Sciences, Beijing 100101, China*

[c] *Bioinformatics Research Group, Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Science, Beijing 100080, China*

[d] *Chinese National Human Genome Center, Beijing 100176, China*

[e] *Department of Theoretical Chemistry, University of Vienna, Währingerstraße 17, A-1090 Wien, Austria*

[f] *The Santa Fe Institute, 1399 Hyde Park Rd., Santa Fe NM 87501*

Abstract

We present a survey for non-coding RNAs and other structured RNA motifs in the genomes of *C. elegans* and *C. briggsae* using the RNAz program. This approach explicitly evaluates comparative sequence information to detect stabilizing selection acting on RNA secondary structure.

We detect 3672 structured RNA motifs, of which only 678 are known ncRNAs or clear homologs of known *C. elegans* ncRNAs. Most of these signals are located in introns or at a distance from known protein-coding genes. With an estimated false positive rate of about 50% and a sensitivity on the order of 50% we estimate that the nematode genomes contain between 3000 and 4000 RNAs with evolutionary conserved secondary structures. Only a small fraction of these belongs to the known RNA classes, including tRNAs, snoRNAs, snRNAs, or microRNAs. A relatively small class of ncRNA candidates is associated with previously observed RNA-specific upstream elements.

Key words: Non-coding RNA, comparative genomics, *Caenorhabditis*.

1 Introduction

Within the last few years, non-coding RNAs have moved from a fringe existence to a central topic in molecular genetics. Starting with the discovery that microRNAs form a generic family of regulators of gene expression, small, non-translated RNAs (ncRNAs) have become a topic of utmost interest in molecular genetics (Bartel and Chen, 2004; Hobert, 2004; Mattick, 2003, 2004; Szymański *et al.*, 2003; Storz *et al.*, 2005). Unlike protein coding genes, ncRNA gene sequences do not exhibit a strong *common* statistical signal that separates them from their genomic context. Individual families of ncRNAs, on the other hand, exhibit evolutionarily very well-conserved secondary structures. Among these are the rRNAs and tRNAs (which are also very well-conserved at the sequence level), as well as both classes of snoRNAs (C/D-box and H/ACA-box snoRNAs), microRNA precursors, the RNA components of RNase P, RNase MRP, SRP, and the five spliceosomal snRNAs (U1, U2, U4, U5, and U6). Structure-based search algorithms such as `ERPIN` (Gautheret and Lambert, 2001), `RNAMotif` (Macke *et al.*, 2001), `Rsearch` (Klein and Eddy, 2003), or `FastR` (Bafna and Zhang, 2004), can thus be used to identify members of these classes in genomic sequences even in the absence of significant sequence homology. These approaches cannot be employed, however, to identify novel RNA families.

The structural conservation of ncRNAs can be understood as a consequence of stabilizing selection acting (predominantly) on the secondary structure. Their sequences, on the other hand, are often highly variable. This results in a substitution pattern that can be utilized to design a general-purpose RNA genefinder based on comparative genomics: The first tool of this type, `qrna` (Rivas and Eddy, 2001), is based upon an SCFG (stochastic context free grammar) method to asses the probability that a pair of aligned sequences evolves under a constraint for preserving a secondary structure. RNAs that are under long-time selection for secondary structure can be expected to have sequences that are more resilient against mutations (Wagner and Stadler, 1999; Nimwegen *et al.*, 1999), which in turn correlates with increased thermodynamic stability of the fold. Indeed, it has been observed that functional RNAs are more stable than the structures formed by randomized sequences (Bonnet *et al.*, 2004; Washietl and Hofacker, 2004; Clote *et al.*, 2005). The program `RNAz` (Washietl *et al.*, 2005a) combines both approaches. It uses a z-score measuring thermodynamic stability of individual sequences and a *structure conservation index* obtained by comparing the folding energies of the individual sequences and the energy of the predicted consensus folding. Both quantities measure different aspects of stabilizing selection acting to preserve RNA structure.

In bacterial genomes, searches for ncRNAs based on the detection of promotor sequences without subsequent ORF were quite successful (Hershberg *et al.*, 2003). In eukaryotes, such a procedure is limited by the diversity and complexity of promotor sequences, the highly variable organization of the genes themselves, and the sheer size of the genomes. The analysis of the flanking sequences of more than 100 experimentally determined ncRNAs in *C. elegans*, however, revealed three distinct upstream motifs common to a number of ncRNA loci both in *C. elegans* and *C. briggsae* (Deng *et al.*, 2005). One coincides with the RNA polymerase-III promoter motif of tRNAs, the second is characteristic for snRNAs, while the third one appears to be specific for a small number of nematode-specific ncRNA transcripts.

A computational survey (Washietl *et al.*, 2005b) for non-coding RNAs with conserved secondary structure in vertebrate, and in particular mammalian, genomes, identified more than 30,000 putative ncRNAs. A similar analysis of the genomes of urochordates (Missal *et al.*, 2005), on the other hand, identified only a few thousand putative structured RNAs, consistent with the hypothesis that ncRNAs form the basis of a complex cellular regulation system that has been vastly expanded in vertebrates (Mattick, 2004; Bartel and Chen, 2004). Here we extend the phylogenetic range of systematic surveys for ncRNAs to nematodes.

2 Methods

2.1 Data sources

The genomic sequence of *C. elegans* was retrieved from the website of the Sanger Institute, i.e., version WS120 of March 2004 [1] , for which a gene and repeat annotation exists at `UCSC genome browser`. For the *C. briggsae* genome (Stein *et al.*, 2004) we used the version `cb25.agp8` of July 2002 [2] . The WormBase gene annotation and the repeat annotation from the `UCSC genome browser` were taken to define non-coding DNA in the *C. elegans* genome.

2.2 Genome-Wide Alignments of Non-Coding DNA

We started with the collection of all contiguous regions of the *C. elegans* genome that are not annotated as either "protein coding in known genes" or as "repetitive elements" in WS120. Putative coding regions predicted by `genscan` or other gene prediction tools were not excluded from this initial data set, which amounts to 61,067,263 bp of the 100,291,769 bp genomic DNA.

[1] `ftp://ftp.sanger.ac.uk/pub/wormbase/FROZEN_RELEASES/WS120/CHROMOSOMES/`

[2] `ftp://hgdownload.cse.ucsc.edu/goldenPath/currentGenomes/Caenorhabditis_briggsae/sanger`

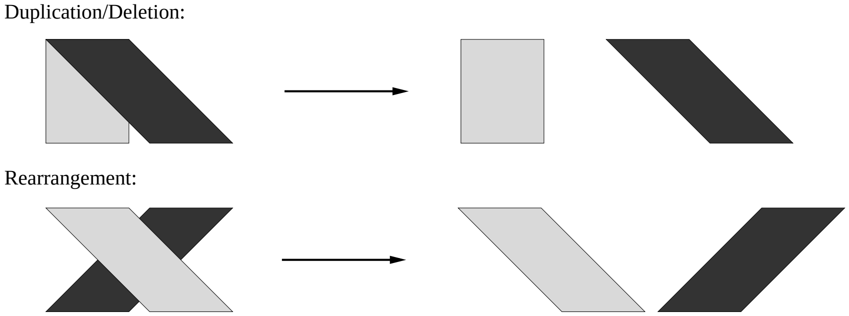


Fig. 1. Local pairwise alignments will lead to an inconsistent global alignment in case of duplication, deletion or rearrangement events. They are combined to a global alignment only if they are consistent.

For each DNA interval, we determined potentially homologous regions in the *C. briggsae* genome by pairwise `blast` (Altschul *et al.*, 1990) searches with $E < 10^{-3}$. Regions separated by only short distances ($\leq$ 30 nt) were combined provided the alignments passed the consistency checks outlined below. Global alignments of the resulting regions were then computed using `clustalw` (Thompson *et al.*, 1994). We obtained pairwise alignments for 13,567,851 bp (13.5%) of the *C. elegans* genome.

Structured RNAs are less conserved in regions without base pair interactions, which might prevent `blast` from extending the sequence alignment into such regions. In order to ensure that a global alignment constitutes a complete ncRNA gene, `blast` hits with short distances between them were combined. But due to rearrangement, deletion, and duplication events during evolution, not all local alignments lead to a consistent global alignment. We therefore employed the following algorithm:

A global alignment is inconsistent if at least one region of sequence A is conserved with at least two regions of sequence B (duplication or deletion) or if at least two distinct regions of sequence A are conserved in different order in sequence B (rearrangement), see Fig. 1. It is useful to construct a graph G_S in the following way: Local alignments are the vertices, and there is an edge between two local alignments if they have a distance less than a threshold value ℓ; in our case $\ell = 30$ nt. The connected components of G_S thus comprise sets of alignments with pairwise short distance; within these, all combinations of consistent, global alignments have to be determined. To this end, one first checks whether each pair x and y of local alignments are consistent, in the sense that they can be derived from the same global alignment. Two further auxiliary graphs G_C and G_I store this consistency information. If x and y are consistent an edge in G_C is introduced, otherwise an edge in G_I is added between x and y. Finally, the graph G_F is constructed by inserting edges between the two nodes x and y if at least one path between x and y exists in

G_C which does not contain pairs of nodes that are inconsistent, i.e., connected by an edge in G_I. Complete subgraphs of G_F correspond to local alignments which can be combined to a consistent global alignment. Only maximal local alignments, i.e., the maximal cliques of G_F, are of interest for our purposes. These can be computed efficiently e.g. by the program *cliquer* (Östergård, 2002). We remark that this approach is similar in spirit to the consistency checking algorithm implemented in the `tracker` algorithm for phylogenetic footprinting (Prohaska *et al.*, 2004).

For some regions, in particular tRNA genes, snRNA genes, and a few other loci we obtained more than one alignment for the same *C. elegans* sequence. This does not constitute a problem for the ncRNA detection, since we obtained essentially identical alignments with different paralogs. Two different alignments of the same reading direction were merged onto the same genomic loci if they overlap to at least 90% in the *C. elegans* genome. All such genomic regions were combined again if they overlap to at least 90% independent of the reading direction of their alignments.

Putative ncRNA clusters in close proximity might still cover a genomic region more than once. Of all merged regions which overlapped more than 20% we discarded all except one leaving us with a unique genomic locus for each ncRNA gene. For each locus we choose the alignment with the maximal `RNAz` classification probability as the representative. Hence, for all statistics reported below, each genomic location is represented in at most one structured RNA candidate.

We used a database system to handle the huge amount of data. We set up a `MySQL 4.1` database server providing sequence information on the *C. elegans* and *C. briggsae* genome including various annotation data. The complete output of the major processing tools `blast` and `RNAz` is stored at the system to allow fast assaying. Currently 18 tables containing up to $1,270,000$ records provide a putative annotation of non-coding RNAs in *C. elegans* and *C. briggsae.*

2.3 Detection of Structured RNA Motifs

The pairwise `clustalw` alignments described above were screened with `RNAz` (Washietl *et al.*, 2005a) to detect regions that are also conserved on the level of RNA secondary structure. Due to computational limitations and restrictions in the training set of the support vector machine (SVM) underlying the `RNAz` program, alignments were scanned by moving a window of length 120 in steps of 50nt. We only scanned alignments of at least 40 nt length, because most known ncRNA families are not shorter than this. The `RNAz` algorithm evaluates the thermodynamic stability of RNA secondary structures (relative to an ensemble of shuffled sequences) and quantifies the evidence for stabilizing selection by comparing the energy of a consensus structure with the ground-state energies of the individual structures. `RNAz` performs the classification by means of a support vector machine that takes into account (1) the length and sequence divergence of the alignment, (2) the number of aligned sequences, (3) the folding energy z-score, and (4) the structure conservation index. A probability estimate $p > p_c$ for the SVM decision value gives a convenient measure to interpret the `RNAz` classification. A value of $p_c = 0.5$ classifies the alignment as non-coding RNA with low significance, whereas $p_c = 0.9$ indicates a high significance for structured RNA. For details we refer to Washietl *et al.* (2005a). For each global alignment, both possible reading directions are considered, because the classification of `RNAz` is based on the thermodynamic stability of the potentially transcribed RNA, which is inherently direction dependent.

2.4 Specificity

In order to estimate the specificity of `RNAz` on the pairwise alignments of non-coding DNA, we repeated the entire screen with shuffled input alignments. The specificity in terms of individual `RNAz` scanning windows is defined as

$$\text{specificity} := \frac{\text{number of shuffled scanning windows with } p \leq p_c}{\text{number of shuffled scanning windows}} \quad .$$

We found that `RNAz` has a specificity of more than 0.96 ($p_c = 0.5$) and 0.98 ($p_c = 0.9$). However, we observe “raw overall false positive rates” of the entire screen of 56% ($p_c = 0.5$) and 41% ($p_c = 0.9$) by comparing the number of genomic regions classified as structured RNA in the true data with the shuffled data set. We define the raw overall false positive rate as

$$\text{raw overall false positive rate} := \frac{\sum_{\{i \in \text{shuffled screen}\}} l_i}{\sum_{\{j \in \text{original screen}\}} l_j} \quad ,$$

where l_i and l_j are the length of the i-th and j-th unique genomic loci classified as ncRNA in the shuffled and original screen, respectively. These raw false-positive rates are, however, dramatic overestimates since we shuffled each alignment independently. Thus, if there are $M > 1$ alignments for a given locus (which is the case for all ncRNA genes that appear in multiple copies in the genome), there are M independently shuffled alignments (Fig. 2). Our procedure, however, counts a locus as a false positive as soon as one of them is misclassified by `RNAz`. In order to correct for this effect we counted each

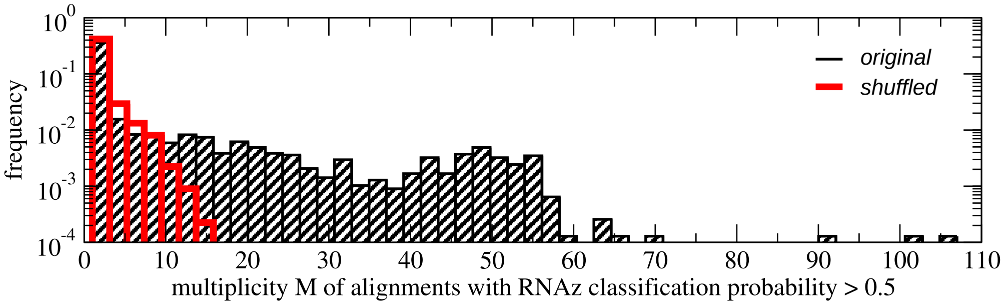


Fig. 2. Distribution of the number M of alignments, classified as structured RNA, mapping to a given genomic locus.

alignment with a weight $1/M$:

$$\text{corrected overall false positive rate} := \frac{\sum_{\{i \in \text{shuffled screen}\}} \frac{l_i}{M_i}}{\sum_{\{j \in \text{original screen}\}} \frac{l_j}{M_j}} ,$$

and obtained corrected false positive rates of 49% ($p_c = 0.5$) and 33% ($p_c = 0.9$), respectively.

Alternatively, we defined an individual false positive rate as

$$\text{individual false positive rate} := \\ \frac{\text{number of shuffled scanning windows with } p > p_c}{\text{number of original scanning windows with } p > p_c} .$$

Based on this definition we obtained the much smaller false positive rates of 10.9% ($p_c = 0.5$) and 5.5% ($p_c = 0.9$). The reason for this difference is that RNAz hits overlap due to the windowing technique. While overlapping windows typically agree on their classification in the true data set, RNAz hits only sparsely cover a misclassified genomic locus in the shuffled dataset. This effect suggests the possibility for further methodological improvements that could increase the specificity of RNAz.

2.5 Estimating the Sensitivity of RNAz

In order to estimate the sensitivity of our screen we compared our data to a recent annotation of non-coding RNAs in *Caenorhabditis elegans* (Stricklin *et al.*, 2005), Tab. 3, and (Deng *et al.*, 2005), Tab. 4. We annotate a putative ncRNA candidate of our screen as known if its genomic locus overlaps to at least 70% with a ncRNA annotated in *C. elegans*, leading to the following definition of sensitivity:

$$s_g := \frac{N}{N_g} \quad .$$

Here, N is the number of unique genomic loci, identified by `RNAz`, which overlap to at least 70% with a known member of a specific ncRNA family found in (Stricklin *et al.*, 2005) or (Deng *et al.*, 2005) and N_g is the entire number of ncRNAs of this family in the genome. The sensitivity of our screen largely depends on the number of ncRNAs which have a conserved primary structure between *C. elegans* and *C. briggsae*. To state how many known ncRNAs in *C. elegans* can be detected by our screen in principle we also report the sensitivity of our alignment procedure defined as

$$s_a := \frac{N}{N_a} \quad ,$$

where N_a is the number of known ncRNAs overlapping to at least 70% with our pairwise alignments scanned with `RNAz`.
While in (Stricklin *et al.*, 2005) the WS130 assembly of *C. elegans* was used, we based our screen on the WS120 assembly, because for WS120 a protein coding gene and a repeat annotation track are provided by UCSC. This allowed us to summarize the results of our survey conveniently as `RNAz` custom-track that can be readily viewed in the `UCSC genome browser`. All `RNAz` hits with classification probability $p_c = 0.5$ were mapped to the WS130 in order to facilitate comparison with the "Wormbook annotation" (Stricklin *et al.*, 2005).

2.6 Upstream Patterns

The putative regulatory motifs considered here were derived from the experimentally determined ncRNAs reported by Deng *et al.* (2005). The 100 bp upstream of these 198 genomic loci were extracted from the genomic DNA sequence and analyzed with the pattern discovery software `meme` (Bailey and Elkan, 1994) with parameters `-dna -nmotifs 10`. Three upstream motifs (UMs) were statistically highly significant and each of them belongs to more than three different RNAs or RNA families; see (Deng *et al.*, 2005) for further details. Most probably, therefore, these elements constitute regulatory (promoter) elements.

The complete *C. elegans* genome was scanned for occurrences of these three UMs using the program `MotifLocator` from the software `INCLUsive` (Thijs *et al.*, 2001). This program uses an adapted position-weight matrix scoring scheme based upon a higher-order background model. The score is computed as the normalized ratio of the motif score and the background score. The threshold value for the score is determined by counting the number of hits of the very abundant UM1 motif with different thresholds. In order to en-

Table 1
Statistics of RNAz ncRNA screens for two different classification probability levels p_c. A comparison of the number of initial blast alignments with the number of ncRNA candidates predicted by RNAz shows that ncRNAs are slightly enriched in introns, while UTR elements are rare.

Genomic context	blast alignments length	Number of ncRNA candidates	
		$p_c = 0.5$	$p_c = 0.9$
intronic	597,128	1235	891
5'UTR	116,193	119	65
3'UTR	128,766	130	69
intergenic	810,989	1221	726
total		3672	2366
length(nt)	13,567,851	432,536	291,499

A ncRNA is classified as "intergenic" if it is at least 1kb away from the closest known protein coding gene in *Caenorhabditis elegans*; a ncRNA is classified as "UTR" if it is located within an interval of GeneBounds track either before the first or after the last coding exon of the gene in question. 54 ncRNAs are annotated as 5'UTR as well as 3'UTR, which might be regulatory elements for polycistronic transcripts (Blumenthal, 2004). All numbers refer to the *C. elegans* genome.

sure that the results do not depend strongly on the software, we compared MotifLocator with PatSearch (Grillo *et al.*, 2003). The threshold score value of 0.8 was chosen since the number of PatSearch hits increases sharply below this value. The results were similar for both softwares, and only the MotifLocator data was used for further analysis.
The motifs identified by the genome wide MotifLocator scan was compared to the RNAz predictions. However, a comprehensive investigation of the upstream regions of the RNAz predictions, unfortunately, is complicated by both the large set of predictions and the fact that RNAz cannot reliably determine the direction and the ends of the putative ncRNAs.

3 Results

3.1 Novel ncRNAs

We detected 3672 structured RNA signals ($p_c = 0.5$) of which 678 correspond to 665 known ncRNAs or clear homologs of known *C. elegans* ncRNAs (Tab. 1,

Table 2
Specificity and false positive rates of the RNAz ncRNA screens for two different classification probability levels p. False positive rates can be estimated in different ways for our screen (see text for details). The estimate for the individual windows that are screened with RNAz appears optimistic, while the estimates for the entire screen are by construction pessimistic.

	False positive rates	
	$p_c = 0.5$	$p_c = 0.9$
individual RNAz hits	10.9%	5.5%
genomic loci (raw)	55.9%	40.9%
genomic loci (corrected)	48.8%	33.2%
Specificity per test	0.96	0.98

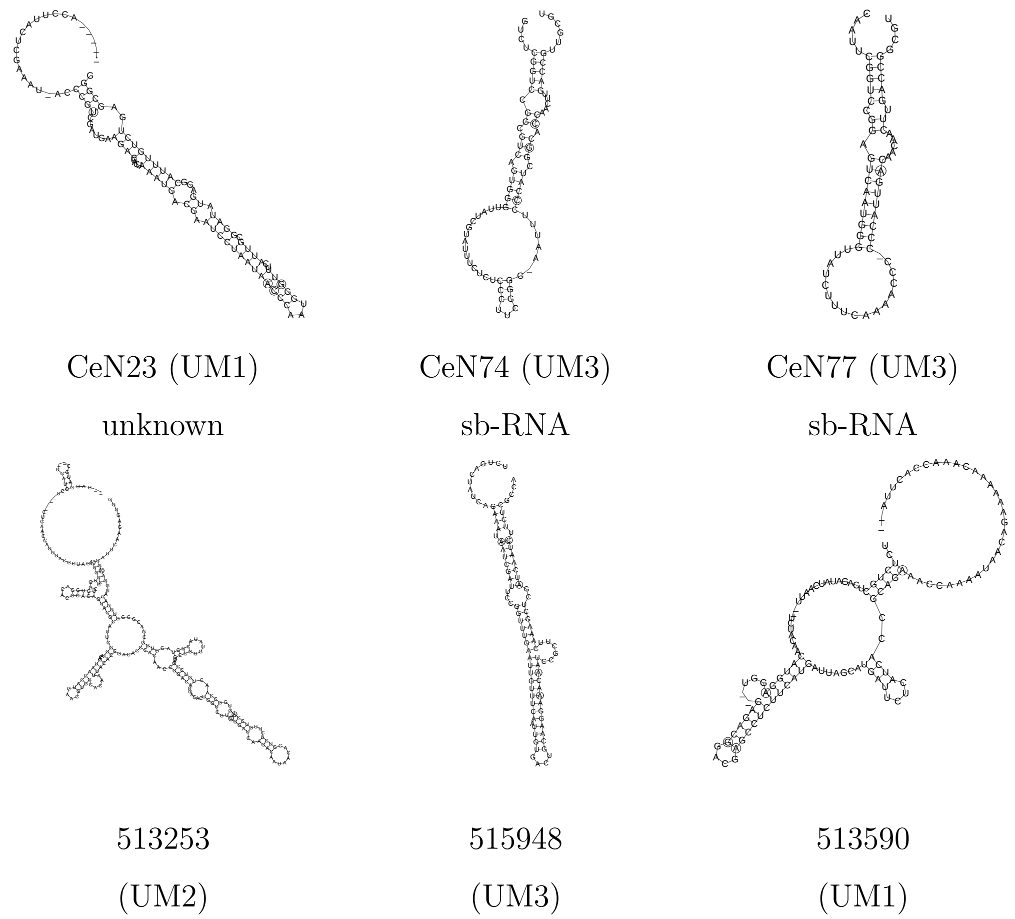
Tab. 3 and Tab. 4). The complete dataset can be accessed as a gff file that is included in the electronic supplement[3]. A few examples are shown in Fig. 3.

Approximately a quarter of the RNAz hits are located in introns, and a comparable number is "intergenic" in the sense that it is located more than 1kb away from any known protein coding gene. Putative RNA structures in untranslated regions (UTRs) of protein coding genes are identified using the GeneBounds track provided at the UCSC Genome Browser. Interestingly, ncRNA candidates have approximately equal densities in intron and intergenic regions, while they are underrepresented by a factor of 10 in UTRs.

3.2 Specificity and Sensitivity of the RNAz screen

Specificity and false positive rates can be estimated by different methods, as outlined in the Methods section. Using the individual alignment windows that are scored by RNAz we observe a false positive rate of less than 11% ($p_c = 0.5$) in a comparison between real and randomly shuffled data. This is probably an optimistic estimate. On the other hand, the false positive rates of the entire screen, corrected for multiple alignments mapping to the same genomic position are about 50% ($p_c = 0.5$) and 33% ($p_c = 0.9$), Tab. 2. As argued above, these are pessimistic estimates. The 3672 ($p_c = 0.5$) and 2366 ($p_c = 0.9$) predicted ncRNAs imply lower bounds between 1600 and 1900 structured RNAs, of which roughly one third (see Tab. 3) are annotated. It follows that we can expect at least roughly 1000 *bona fide* novel ncRNAs and structured RNA elements in our dataset.

[3] URL: http://www.bioinf.uni-leipzig.de/Publications/SUPPLEMENTS/05-023/



CeN23 (UM1)
unknown

CeN74 (UM3)
sb-RNA

CeN77 (UM3)
sb-RNA

513253
(UM2)

515948
(UM3)

513590
(UM1)

Fig. 3. Examples of ncRNAs in *C. elegans*. The top row shows predicted consensus structures for three ncRNAs experimentally verified by Deng *et al.* (2005), the associated upstream motif is listed in parentheses. SbRNAs (stem-bulge RNAs) are a set of conserved nematode ncRNAs showing two conserved motifs located at the 5' and 3'end of the transcript, which together form an imperfect stem with a characteristic bulge. The second row shows three `RNAz` predictions that are associated with one of the upstream motifs reported by Deng *et al.* (2005). Small circles indicate consistent and compensatory mutations, respectively.

An accurate estimate of the overall sensitivity of an RNA gene finding approach is hard to derive since comprehensive annotations are available only for a few "classical" families of ncRNAs. In the following we briefly outline our results for the major ncRNA families; see also Tab. 3.

To annotate the RNA classes below we mapped the annotation in (Stricklin *et al.*, 2005), which is given in WS130 coordinates, to the *C. elegans* assembly WS120, from which we derived our sequences. An overlap of at least 70% is required for a ncRNA candidate to be annotated.

tRNAs. Of about 591 known tRNAs in *C.elegans*, we identified 509 ($p_c = 0.5$) and 465 ($p_c = 0.9$) in our screen. Only 70 of 1072 tRNA pseudogenes are found in our global pairwise alignments of which 50 at $p_c = 0.5$ and 44 at $p_c = 0.9$ were detected by `RNAz`.

rRNAs. About three 18S, one 23S, one 26S, fifteen 5S and two 5.8S are known in *C. elegans.* We recovered all 18S rRNAs in both chromosome I and in the mitochondrial DNA. The mitochondrial 23S rRNA appears as two separated `RNAz` hits. The repeat unit of 26S rRNAs on chromosome I is also detected as a series of ten separated `RNAz` hits. One single copy of the 5S rRNA in chromosome V (with an overlapping constraint of at least 60%) was detected by our screen, but none of the 5.8S rRNAs. Both 5.8S rRNAs are not conserved in *C. briggsae* (`blast` cutoff of $E < 10^{-3}$) and hence are not identifiable by our approach. Whereas the fifteen known 5S rRNA loci are well conserved in *C. briggsae* (`blast` cutoff of $E < 10^{-3}$), their sequence similarities (96% to 100%) are beyond the favourable values for `RNAz`. In such alignments no covariance information to predict a reliable consensus secondary structure is given and the high degree of structure conservation, resulting from almost identical sequences, is not significant.

miRNAs. In *C. elegans*, 117 miRNAs are annotated in (Stricklin *et al.*, 2005). Of these, 54 are conserved with *C. briggsae* at a `blast` cutoff of $E < 10^{-3}$. While the mature miRNA sequence is easily detected by `blast`, we failed to detect the precursor stem loop of some miRNAs in our pairwise alignments. Indeed, only 40 of the 54 conserved miRNAs overlap to at least 70% with pairwise alignments longer than 40 nt. Only those were scanned by `RNAz` and were therefore in principle identifiable by our screen. We detected 34 of these 40 miRNA precursor genes at both $p_c = 0.5$ and $p_c = 0.9$.

snoRNAs. Of the 31 known small nucleolar RNA genes we found 13 at the $p_c = 0.5$ level and only 9 at $p_c = 0.9$. Fifteen of the annotated snoRNAs are experimentally verified. We detected 10 of these at $p_c = 0.5$ and 9 at $p_c = 0.9$. This amounts to a sensitivity of 0.66 ($p_c = 0.5$) and 0.60 ($p_c = 0.9$), respectively. Of the 16 annotated snoRNAs which are not experimentally verified, we could only identify 3 at $p_c = 0.5$ and none at $p_c = 0.9$.

RNaseP RNA. The one known copy of RNase P RNA was detected by our screen with a classification probability $p_c = 0.99$. In contrast we do not find a RNAse MRP RNA. If *Caenorhabditis* species have a RNAse MRP RNA it appears to be highly divergent from other species. A recent specific search for this ncRNA did not detect candidates in either *C. elegans* or *C. briggsae* (Piccinelli *et al.*, 2005).

Spliced Leader RNAs. The first form, SL1 RNA, occurs in 10 copies in a tandem repeat in chromosome V, whereas the second form, SL2 RNA, is

Table 3
Sensitivity of RNAz-detected ncRNAs based on known ncRNA annotations from the Wormbook (Stricklin *et al.*, 2005). We compare the numbers N_g of genes known in the genome (2nd column) with those contained in our input alignments (N_a), and those classified as structured RNAs by RNAz (N) at two different classification probability levels. In addition, sensitivities are listed as fraction s_g of known genomic sequences, and as fraction s_a of known sequences contained in the input alignments (given in brackets).

		known in	in *C.el./C.br*		RNAz					
		genome	alignment		$p_c = 0.5$			$p_c = 0.9$		
		N_g	N_a	s_g	N	s_g	s_a	N	s_g	s_a
tRNA	functional	591	584	0.98	509	0.86	[0.87]	465	0.78	[0.79]
	pseudogene	1072	70		50			44		
miRNA		117	40	0.34	34	0.29	[0.85]	34	0.29	[0.85]
snoRNA		31	26	0.84	13	0.41	[0.50]	9	0.29	[0.35]
snRNA	spliceosomal	72	72	1.00	54	0.75	[0.75]	47	0.65	[0.65]
	spliced leader	30	26	0.87	26	0.87	[1.00]	26	0.87	[1.00]
rRNA		22	20	0.9	5	0.22	[0.25]	4	0.18	[0.2]

The sensitivity of the miRNA genes refers to the 54 miRNA loci conserved in *C. briggsae*. For all other ncRNA classes the sensitivity values refer to the number of the known genomic loci in *C. elegans*. Known ncRNA genes are counted to be in our alignments if they overlap to at least 70% with a global alignment.
Sensitivities are also reported relative to the *C. elegans* genome.

found in 20 variants. At both $p_c = 0.5$ and $p_c = 0.9$ we found 10 regions in chromosome V which overlap with the 10 known SL1 RNA genes and 16 variants of the SL2 gene.

Spliceosomal RNAs. 12 U1, 19 U2, 5 U4, 13 U5, and 23 U6 spliceosomal RNA genes are known in *C. elegans*. At $p_c = 0.5$, we could identify all the known U1, U2, U4, and U5 genes and 5 of the U6 loci. At $p_c = 0.9$ we missed 2 U4, 1 U5 and 4 U6 RNA genes.

A recent experimental screen for ncRNAs in *C. elegans* (Deng *et al.*, 2005) described 161 ncRNA transcripts mapping to 198 genomic loci, of which 100 transcripts at 101 loci were unknown before this study. A subset of 69 distinct sequences are putative snoRNA-like transcripts and 31 are functionally unassigned. This set of 100 “novel” ncRNAs provides us at least with a rough estimate on how our comparative genomics approach performed beyond the realm of the “classical ncRNAs”. Since tRNAs and rRNAs (which form a

Table 4
Comparison of the RNAz results with the experimental small RNA screen (Deng *et al.*, 2005). All numbers refer to genomic locations in the *C. elegans* genome. Columns have the same meaning as in the previous Tab. 3.

Type	Deng 2005 ncRNA loci	*C.el*/*C.br.* alignments		RNAz					
				$p_c = 0.5$			$p_c = 0.9$		
	N_g	N_a	s_g	N	s_g	s_a	N	s_g	s_a
in Wormbook	97	90	0.93	63	0.64	[0.70]	55	0.56	[0.61]
H/ACA snoRNA	41	31	0.76	11	0.26	[0.35]	9	0.21	[0.29]
CD snoRNA	28	19	0.68	3	0.10	[0.15]	2	0.07	[0.10]
sb RNA	9	3	0.33	2	0.22	[0.66]	2	0.22	[0.66]
snl RNA	8	3	0.38	3	0.37	[1.00]	2	0.25	[0.66]
unknown	14	14	1.00	4	0.28	[0.28]	2	0.14	[0.14]
all novel	101	70	0.69	23	0.23	[0.33]	17	0.17	[0.24]
Total	198	160	0.81	86	0.43	[0.53]	72	0.36	[0.45]

The coordinates given by Stricklin *et al.* (2005) and Deng *et al.* (2005) are mapped to WS120, the coordinates of our ncRNA candidates. Annotations overlapping at least 70% with RNAz hits are counted as the same ncRNA gene. Sensitivities are given relative to both the genomic loci, and relative to the loci that are contained in our alignments (in square brackets).

substantial fraction of the known ncRNAs) are among the evolutionarily best conserved genes, it is to be expected that they are easier to find and recognize as structured RNAs than most other ncRNAs. Indeed, the sensitivity of RNAz on this dataset is significantly lower: we recovered only 23 of the 101 non-Wormbook loci, Tab. 4.

Tab. 3 and 4 show that the sensitivity of our screen can be understood in terms of two effects: the classification accuracy of RNAz, and the probability that the corresponding genomic region is sufficiently conserved to yield a blast-based alignment. With the exception of the annotated microRNAs, which contain a large number of species-specific sequences annotated as “tiny non-coding RNAs” by Ambros *et al.* (2003), more than 80% of the well-conserved classical ncRNAs (tRNAs, rRNAs, RNase P and MRP, pre-miRNAs, snRNAs) are contained in alignments, while the fraction is smaller for snoRNAs. In the case of snoRNAs, the Wormbook annotation seems to have a bias towards the few snoRNAs with rather well-conserved sequences.

The sensitivity of RNAz strongly depends on the RNA class. It is typically on

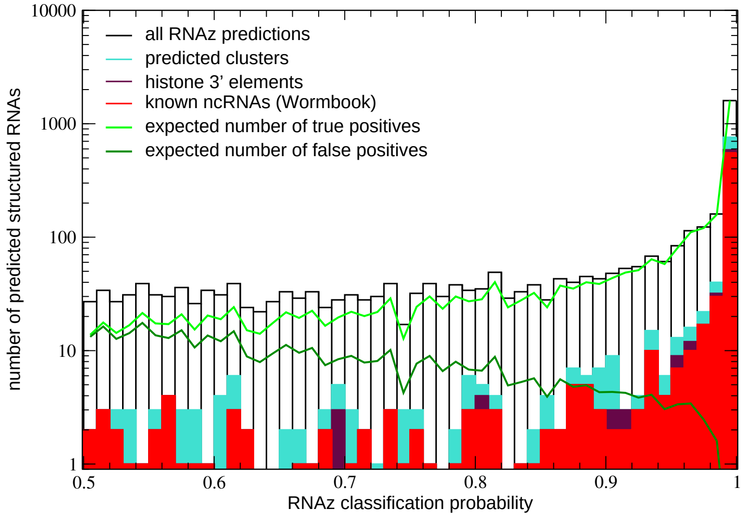


Fig. 4. Distribution of classification probabilities p among RNAz predictions. Colors indicate the fractions of known ncRNAs, predicted histone elements, and predicted families with two or more homologous in each histogram bar.

the order of 80%, with the notable exception of snoRNAs, which are notoriously hard to recognize based on sequence alignments (Washietl *et al.*, 2005b). For this class we have a sensitivity of one-third to one-half. The low sensitivity for rRNAs is due to the high degree of conservation of the 5S rRNAs between *C. elegans* and *C. briggsae*, which makes it impossible for RNAz to make a significant decision, because the global alignments lack any covariance information. All other classes of rRNAs, with the exception of 5.8S rRNA which is not conserved in *C. briggsae*, were successfully identified as structured ncRNAs. We estimate that the sensitivities observed on the data set from (Deng *et al.*, 2005) is probably a plausible order of magnitude of the overall sensitivity of our screen, that is approximately 25-50%.

The support vector machine underlying the RNAz program classified the overwhelming majority of known ncRNAs as “structured RNA” with classification probabilities $p_c = 0.9$, Fig. 4. Nevertheless, a significant number of true positives is identified with small values of p_c, indicating that a cutoff at a much higher value of p_c, than 0.5, would significantly decrease the sensitivity.

The distribution of classification probabilities p also provides us with an independent way of estimating the false positive rate, yielding a value of about 11%, in agreement with the observed false positive rate for individual RNAz hits. The much less favorable false positive rate of 49% for the entire screen has it roots in the overlapping RNAz hits and the fact that our procedure by construction systematically overestimates the false positive rate.

Table 5
Three upstream motifs discovered by Deng *et al.* (2005) are associated with `RNAz`-predicted ncRNAs ($p_c = 0.5$). We separately show the association with tRNAs (mostly UM2), the experimentally verified ncRNAs described by Deng *et al.* (2005), other known ncRNAs according to the Wormbook annotation (Stricklin *et al.*, 2005), and novel candidates. In the latter case we distinguish between motifs that overlap `RNAz` hits and those in a close distance upstream of the `RNAz` signal. The number of unique hits can be less than the sum of the columns 2, 3, and 4 if a `RNAz`-predicted ncRNA is associated to more than one putative promoter sequence.

Hit type	UM1	UM2	UM3	#unique hits
tRNAs	0	391	0	391
(Deng *et al.*, 2005)	55	6	4	63
other known	18	2	0	20
unknown overlapping	1	3	1	4
unknown ≤ 500	16	11	2	28
total	90	413	7	506
predicted (`MotifLocator`)	2182	2390	92	4664

3.3 RNA-Specific Promotors

Deng *et al.* (2005) have identified three putative RNA-specific promotor sequences, denoted by UM1, UM2, and UM3.

Upstream motif UM1 was found at the loci of both snRNAs and a number of other known and novel *C. elegans* ncRNAs, and includes the *C. elegans* proximal sequence element (PSE) characteristic for spliceosomal snRNAs (Thomas *et al.*, 1990; Hernandez, 2001).

UM2 was mainly found upstream of snoRNA genes. However, the motif also bears a strong resemblance to the internal tRNA promoter, and indeed 1135 UM2 elements overlap 391 of 591 tRNA and 745 of 1072 tRNA-pseudo-genes according to the annotation of (Stricklin *et al.*, 2005).

The third motif, UM3, was only found at the loci of seven transcripts, of which five are functionally unassigned, one is annotated as U6 snRNA, and another as RNAse P.

A hand-curated list of ncRNA candidates from the `RNAz` screen that are associated with one or more of the three upstream motifs was produced in the following way: We combined the positions from the annotation in (Stricklin *et al.*, 2005), of the predicted upstream motifs from (Deng *et al.*, 2005), from our RNAz screen, and the transcripts reported in (Deng *et al.*, 2005). The positions were sorted numerically and combined into clusters if the distance of consecutive annotations was at most 500nt. Tab. 5 summarizes the annotation of the putative ncRNAs that are associated with one of the three promotor motifs. Of the 536 initial hits we retained 506, in the remaining cases the promotor was directed away from the RNAz hit. As expected, the majority of the overlaps are tRNA/UM2 combinations.

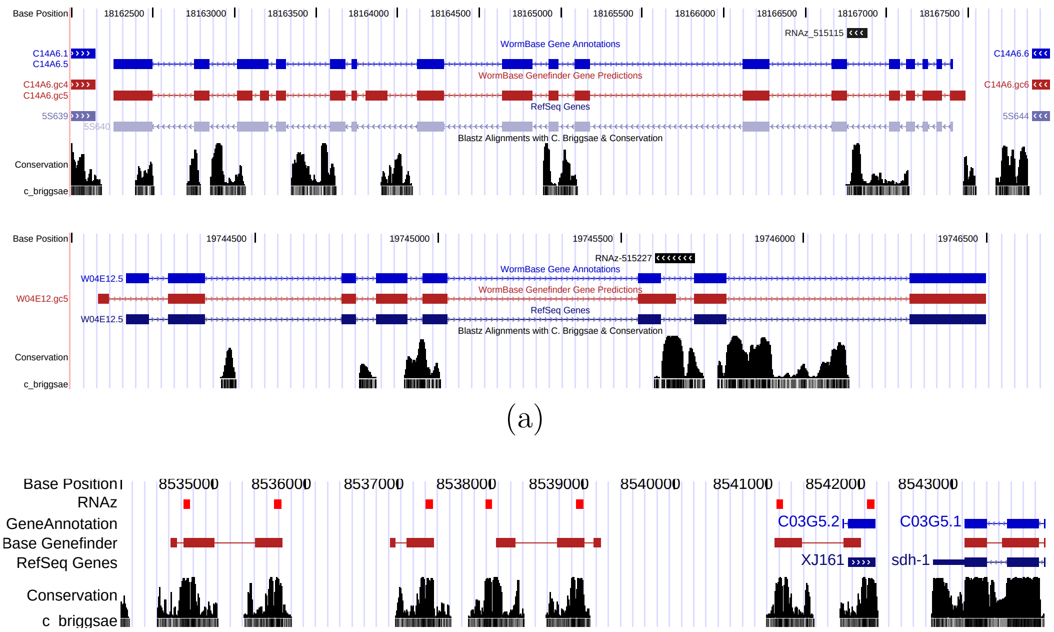
3.4 Intronic ncRNAs

A large fraction of our ncRNA candidates are located in introns. Interesting examples are *RNAz-515115* and *RNAz-515227*, which are located in introns of the putative protein coding genes *C14A6.5* and *W04E12.5*, respectively. Both genes do not have annotated homologs in *C. briggsae* but their intronic sequences are fairly well-conserved in *C. briggsae*, Fig. 5a. Such a structure is reminiscent of "host genes" whose only purpose is to carry snoRNAs in their introns (Tycowski *et al.*, 1996; Tycowski and Steitz, 2001; Bachellerie *et al.*, 2002). However, using snoscan (Lowe and Eddy, 1999) to test for C/D box snoRNAs and checking the secondary consensus structure for two hairpins typic for H/ACA box snoRNAs, failed to produce evidence that *RNAz-515115* and *RNAz-515227* may be snoRNAs. The consensus structures, predicted by RNAz, are well conserved and stable hairpins. We therefore presume that both ncRNA candidates may be miRNAs, which is supported by (Rodriguez *et al.*, 2004) where it is shown that also miRNAs occur in "host genes".

3.5 Multi-copy structured RNAs

Clustering the RNAz hits using blastclust with a minimum overlap of 50% and a minimum sequence identity of 50% yields 148 clusters containing a total of 916 RNAz signals and 2756 individual sequences. Most of the sequences in these clusters are known tRNAs, snRNAs, and other ncRNAs for which an unambiguous annotation is available (725 sequences in 134 clusters).

The largest remaining group is associated with histone genes. An initial analysis of the blastclust results of this group gave 6 clusters containing a total of 36 sequences that mapped to various annotated histone genes in *C. elegans*. The consensus of these sequences was then compared to the complete *C. elegans* genome, yielding a total of 47 RNAz hits. The motif appears in a region that is annotated as "a consensus sequence thought to contain a putative U7 snRNA" in two GenBank entries of *C. elegans* histone genes ***X15633*** and ***X15634*** (Roberts *et al.*, 1989). The U7 snRNA is part of the machinery for processing histone mRNAs (see (Dominski and Marzluff, 1999) for a review)



(a)

(b)

Fig. 5. (a) Location of *RNAz-515115* in the gene *C14A6.5* and *RNAz-515227* in the gene *W04E12.5*. The exonic sequences of those genes are not conserved in *C. briggsae*, while most of the intronic sequence is rather well conserved.
(b) Genomic context of 7 of the eight members of a cluster of related `RNAz` hits containing *RNAz-515800*. This cluster is localized at the *C. elegans* X-chromosome. Some of the cluster members overlap with a predicted protein-coding gene.

but so far has not been verified directly in nematodes. We checked for Sm protein binding site, HDE binding site and the snRNA-like promoter element UM1, however with negative result. We conclude that the histone-3'-motif corresponds to the hairpin motif found in histone mRNAs of other species.

Seventeen additional `blastclust` clusters contain more than 2 genomic loci. One of these cluster appears to be associated with the multigene family of MSPs (major sperm proteins), while the 16 other clusters are not related to annotated protein-coding genes. We extracted well-conserved consensus sequences for those elements and then performed a `blast` search against the database of `RNAz` hits with $E < 10^{-10}$. In total, we find 216 sequences in 127 `blastclust` clusters that match one of these consensus sequences. Of the remaining hits, 53 appear twice and 2577 are single sequence motifs.

A few of these `blastclust` clusters are localized in one or a few narrow genomic regions, an example is shown in Fig. 5b. Consensus sequences of these multi-copy sequences are given in the electronic supplement. In many cases, there is evidence for some form of concerted evolution since the *C. elegans* loci are more similar among themselves than compared to the homologous *C.*

briggsae sequences. One of these families forms very stable hairpins and hence might be microRNA precursors.

To date, no telomerase RNA has been reported for *C. elegans* (Jones *et al.*, 2001; Stein *et al.*, 2004; Stricklin *et al.*, 2005), although — in contrast to Drosophila (Melnikova and P, 2005) — this species has a "normal" telomeric repeat sequence. A putative telomerase reverse transcriptase (TERT) was identified by Malik *et al.* (2000), which shows several atypical features, suggesting that a unique mechanism of telomere extension may have developed in the Caenorhabditis lineage. We therefore further investigated the `RNAz` hits containing the one-and-a-half repeat of the telomeric template that is characteristic for telomerase RNA (Jones *et al.*, 2001), here CCTAAGCCTTAA. The set of 16 candidates (excluding intronic and UTR elements) does not contain the two putative telomerase RNAs *tts-1* and *tts-2* transcripts discussed by Jones *et al.* (2001): The first is not conserved at sequence level, the second is not classified as structured RNA by `RNAz`. We checked, using `pknotsRG` (Reeder and Giegerich, 2004), for a locally stable pseudoknot domain immediately downstream of the template sequence, that is typically observed in vertebrate, ciliates and yeast telomerase RNAs (Chen *et al.*, 2000; Lin *et al.*, 2001; Chen and Greider, 2005). Inspection of the resulting 5 candidates showed, however, that the *C. briggsae* sequences in these alignments contain longer repetitive stretches of the (reverse complement of the) telomeric repeat sequence, suggesting that they may be false positives arising from aligning the *C. elegans* sequence with repetitive sequences from *C. briggsae*. Consequently, our survey did not detect a plausible candidate for telomerase RNA with a conserved secondary structure.

3.6 Novel MicroRNA Candidates

Possible novel microRNA precursors can be identified by a rather crude filtering procedure from the set of all `RNAz` hits. All `RNAz` hits are realigned with their homologs in *C. briggsae* and those without a conserved hairpin structure are discarded. The conserved hairpin structure is extracted and the restricted alignment is scanned with `RNAz`. It is accepted as a pre-miRNA candidate provided (1) it forms a stem-loop structure with a total length between 40 and 130 and (2) its z-score is below $z = -3.0$. This threshold value was identified by assessing pairwise alignments of random chosen homologous sequences from the `Rfam` database (Missal *et al.*, 2005). Fig. 6 summarizes the comparison of the filtered `RNAz` hits with the candidate set proposed by Grad *et al.* (2003) and the set of known miRNAs of *C. elegans* (Stricklin *et al.*, 2005).

We expect that this simple filter has a rather large false positive rate. In addition, its sensitivity is rather limited, recovering only 22 of the 34 known

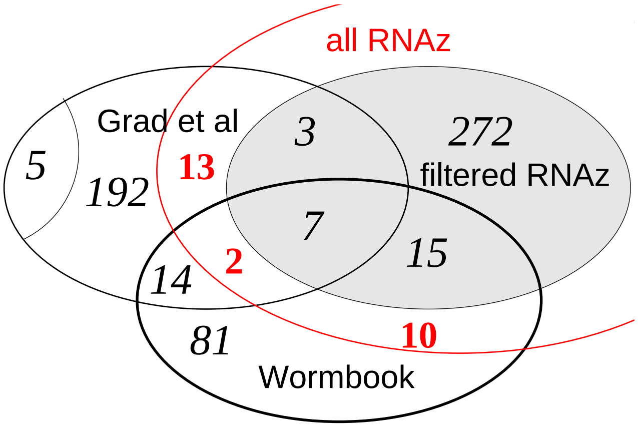


Fig. 6. Comparison of microRNA candidates derived from the `RNAz` screen with the 222 microRNA candidates from (Grad *et al.*, 2003) (4 of these could not be mapped to the WS120 genome assembly and 1 is apparently a repetitive element) and the 117 known miRNAs listed by Stricklin *et al.* (2005). Numbers in red refer to `RNAz` hits that did not pass the simple filter described in the text.

microRNA precursors detected by our screen (Tab. 3). A more sophisticated post-processing using e.g. `miRscan` (Lim *et al.*, 2003) should provide better results; this program, however, is only available as a web-service and hence not suitable to screen the entire set of thousands of `RNAz` predictions.

4 Discussion

The systematic comparison of the genomic DNA of *C. elegans* and *C. briggsae* reveals evidence for a large number of structured RNA motifs. Most are located either within introns or relatively far away from known protein coding regions. This strongly suggests that the majority of these signals are *bona fide* noncoding RNAs. The comparable density of signals in introns and intergenic regions, and the very sparse occurrence of signals in UTRs also tally well with a recent experimental study of *C. elegans* ncRNAs, in which 56% of 198 loci were found overlapping an intron versus 42% in intergenic regions, and only very few loci found in UTRs (Deng *et al.*, 2005). The argument for `RNAz` signals representing actual ncRNA loci is further supported by the fact that some subclasses of both intronic and intergenic ncRNAs are associated with upstream motifs that appear to be characteristic for *C. elegans* ncRNAs.

With an estimated sensitivity of around 50% we therefore predict the total number of structured RNA motifs at 3000-4000, comprising about 1Mb of the genome. We emphasize that our survey is based on the `RNAz` program (Washietl *et al.*, 2005a), which is based on both primary sequence conservation and secondary structure conservation. Both are factors which may reduce the sensitivity of our screen, because very much of the recently detected non-coding transcription is poorly conserved between relatively close species (Wang *et al.*, 2004; Hyashizaki, 2004) and RNAs which might perform their function without the need for a well-defined structure, for example anti-sense transcripts (FANTOM Consortium *et al.*, 2005), are not detectable by our method. This could also explain the small fraction of `RNAz` hits that are associated with upstream motifs. Nevertheless, estimates based on intron conservation and conserved upstream motifs have arrived at figures in the range 1600 to 4100 different ncRNAs *C. elegans* (Deng *et al.*, 2005), thus lending support to our estimate for structured RNA motifs.

These numbers have to be compared with estimates for the ncRNA content in other genomes. An `RNAz` survey based on the most conserved parts of the vertebrate genomes estimates that the ncRNA content of mammalian genomes is comparable to their protein coding genes (Washietl *et al.*, 2005b), and hence at least an order magnitude larger than in nematodes. In contrast, the predicted number of structured RNAs in the urochordate *Ciona intestinalis* is comparable to our results for the nematodes (Missal *et al.*, 2005). This indicates that higher vertebrates have dramatically expanded their ncRNA inventory relative to their complement of protein coding genes. This is consistent with the assumption that the function of the ncRNAs is primarily regulatory (Mattick, 2003, 2004).

The partial analysis of the predicted *C. elegans* ncRNAs highlights important open problems in computational RNomics. With the exception of rRNAs and tRNAs, efficient and reliable tools for classifying ncRNAs are not available. Recent advances in snoRNA detection (Schattner *et al.*, 2005) still require explicit knowledge of the modification targets and hence cannot correctly classify snoRNAs with non-canonical targets such as mRNAs. Even for microRNAs, reliable classification tools that could be used for genome-wide studies are not available. Indeed, for the majority of predicted structured RNAs we have no annotation at all, and the overwhelming majority of them has no homologs outside the nematodes that could be detected unambiguously by means of sequence comparison.

In the near future, several additional nematode genomes will become available, including both distant species with a parasitic lifestyle such as *Brugia malayi* (Ghedin *et al.*, 2004) and close relatives such as *C. remanei*[4]. These additions will bring the total number of sequenced nematode genomes to a total of ten. A denser taxon coverage of nematodes will undoubtedly also increase the specificity of the non-coding RNA annotation in this phylum. Of particular interest are the close relatives within the Caenorhabditis taxon, because for

[4] `http://www.genome.wustl.edu/projects/cremanei/`

these the sequence similarity is sufficient to obtain reasonable-quality genomic alignments. For example, 3066 of the 3672 `RNAz` predicts show significant sequence homology (`blast` with $E < 10^{-5}$) in the current assembly[5] of the *C. remanei* genome. Of these, 1872 are classified as structured RNAs using `RNAz` on multiple alignments composed of the sequences from all three species.
In contrast, only 694 hits are found in a comparison with the *Brugia malayi* genome[6]. More than 90% of these can be identified as tRNAs and other well-known ncRNAs. Since both sensitivity and specificity of comparative genomics approaches such as `RNAz` increase with the amount of available data, a reliable annotation of nematode structured RNAs is at least within reach.

Acknowledgments

The study reported here was conceived by the last two authors during the workshop "Network: From Biology To Theory" held at the Sino-German Center for Research Promotion in Beijing in April 2005. Significant practical assistance from Liu Beifang and fruitful discussions with Ivo L. Hofacker and Stefan Washietl are gratefully acknowledged. This work is supported by the German *DFG* Bioinformatics Initiative, grant BIZ-6/1-2, and the Chinese National Key Basic Research & Development Program 973, grants 2002CB713805 and 2003CB7159.

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ, 1990. Basic local alignment search tool. J Mol Biol 215:403–410.

Ambros V, Lee RC, Lavanway A, Williams PT, Jewell D, 2003. MicroRNAs and other tiny endogenous RNAs in *C. elegans*. Curr Biol 13:807–818.

Bachellerie JP, Cavaillé J, Hüttenhofer A, 2002. The expanding snoRNA world. Biochimie 84:775–790.

Bafna V, Zhang S, 2004. FastR: Fast database search tool for non-coding RNA. Proc IEEE Comp Systems Bioinformatics Conference (pp. 52–61).

Bailey TL, Elkan C, 1994. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. In: Proceedings of the Second

[5] `ftp://genome.wustl.edu/pub/seqmgr/remanei/pcap/remanei_041227/`

[6] `http://www.tigr.org/`. "Preliminary sequence data for *B. malayi* is deposited regularly into the GSS division of GenBank. The Sequencing effort is part of the International Brugia Genome Sequencing Project and is supported by an award from the National Institute of Allergy and Infectious Diseases, National Institutes of Health."

International Conference on Intelligent Systems for Molecular Biology, (pp. 28–36). Menlo Park, CA: AAAI Press.

Bartel DP, Chen CZ, 2004. Micromanagers of gene expression: the potentially wide-spread influence of metazoan microRNAs. Nature Genetics 5:396–400.

Blumenthal T, 2004. Operons in eukaryotes. Briefings in Functional Genomics and Proteomics 3:1–13.

Bonnet E, Wuyts J, Rouzé P, Van de Peer Y, 2004. Evidence that microRNA precursors, unlike other non-coding RNAs, have lower folding free energies than random sequences. Bioinformatics 20:2911–2917.

Chen JL, Blasco MA, Greider CW, 2000. Secondary structure of vertebrate telomerase RNA. Cell 100:503–513.

Chen JL, Greider CW, 2005. Functional analysis of the pseudoknot structure in human telomerase rna. PNAS 102:8080–8085.

Clote P, Ferré F, Kranakis E, Krizanc D, 2005. Structural RNA has lower folding energy than random RNA of the same dinucleotide frequency. RNA 11:578–591.

Deng W, Zhu X, Skogerbø G, Zhao Y, Fu Z, Wang Y, He Housheng Cai L, Sun H, Liu C, Li BL, Bai B, Wang J, Cui Y, Jai D, Wang Y, Du D, Chen R, 2005. Organisation of the *Caenorhabditis elegans* small noncoding transcriptome: genomic features, biogenesis and expression. Genome Res In press.

Dominski Z, Marzluff WF, 1999. Formation of the 3'-end of histone mRNA. Gene 239:1–14.

FANTOM Consortium, RIKEN Genome Exploration Research Group, Genome Science Group, 2005. The transcriptional landscape of the mammalian genome. Science 309:1559–1563.

Gautheret D, Lambert A, 2001. Direct RNA motif definition and identification from multiple sequence alignments using secondary structure profiles. J Mol Biol 313:1003–1011.

Ghedin E, Wang S, Foster JM, Slatko BE, 2004. First sequenced genome of a parasitic nematode. Trends Parasitol 20:151–153.

Grad Y, Aach J, Hayes GD, Reinhart BJ, Church GM, Ruvkun G, Kim J, 2003. Computational and experimental identification of *C. elegans* microRNAs. Mol Cell 11:1253–1263.

Grillo G, Licciulli F, Liuni S, Sbisa E, Pesole G, 2003. `PatSearch`: A program for the detection of patterns and structural motifs in nucleotide sequences. Nucleic Acids Res 31:3608–3612.

Hernandez N, 2001. Small nuclear RNA genes: a model system to study fundamental mechanisms of transcription. J Biol Chem 276:26733–26736.

Hershberg R, Altuvia S, Margalit H, 2003. A survey of small RNA-encoding genes in *Escherichia coli.* Nucl Acids Res 31:1813–1820.

Hobert O, 2004. Common logic of transcription factor and microRNA action. Trends Biochem Sci 29:462–468.

Hyashizaki Y, 2004. Mouse transctiptome: Neutral evolution of 'non-coding' complementary DNAs (reply). Nature 431:2.

Jones SJM, Riddle DL, Pouzyrev AT, Velculescu VE, Hillier L, Eddy SR, Stricklin SL, Baillie DL, Waterston R, Marra MA, 2001. Changes in gene expression associated with developmental arrest and longevity in *Caenorhabditis elegans*. Genome Research 11:1346–1352.

Klein RJ, Eddy SR, 2003. RSEARCH: Finding homologs of single structured RNA sequences. BMC Bioinformatics 4:1471–2105.

Lim L, Lau N, Weinstein E, Abdelhakim A, Yekta S, Rhoades M, Burge C, Bartel P, 2003. The microRNAs of caenorhabditis elegans. Genes Development 17:991–1008.

Lin J, Ly H, Hussain A, Abraham M, Pearl S, Tzfati Y, Parslow TG, Blackburn EH, 2001. A universal telomerase RNA core structure includes structures motifs required for binding the telomerase reverse transcriptase protein. PNAS 101:14713–14718.

Lowe TM, Eddy SE, 1999. A computational screen for methylation guide snoRNAs in yeast. Science 283:1168–71.

Macke TJ, Ecker DJ, Gutell RR, Gautheret D, Case DA, Sampath R, 2001. RNAMotif, an RNA secondary structure definition and search algorithm. Nucl Acids Res 29:4724–4735.

Malik HS, Burke WD, Eickbush TH, 2000. Putative telomerase catalytic subunits from *Giardia lamblia* and *Caenorhabditis elegans*. Gene 251:101–108.

Mattick JS, 2003. Challenging the dogma: the hidden layer of non-protein-coding RNAs in complex organisms. Bioessays 25:930–939.

Mattick JS, 2004. RNA regulation: a new genetics? Nature Genetics 5:316–323.

Melnikova L, P G, 2005. *Drosophila*: the non-telomerase alternative. Chromosome Research 13:431–441.

Missal K, Rose D, Stadler PF, 2005. Non-coding RNAs in *Ciona intestinalis*. Bioinformatics 21(S2):i77–i78.

Nimwegen vE, Crutchfield JP, Huynen M, 1999. Neutral evolution of mutational robustness. Proc Natl Acad Sci USA 96:9716–9720.

Östergård PRJ, 2002. A fast algorithm for the maximum clique problem. Discr Appl Math 120:195–205. Software: `http://users.tkk.fi/~pat/cliquer.html`.

Piccinelli P, Rosenblad MA, Samuelsson T, 2005. Identification and analysis fo ribonuclease P and MRP RNA in a broad range of eukaryotes. Nucleic Acids Res 33:4485–95.

Prohaska S, Fried C, Flamm C, Wagner G, Stadler PF, 2004. Surveying phylogenetic footprints in large gene clusters: Applications to Hox cluster duplications. Mol Phylog Evol 31:581–604.

Reeder J, Giegerich R, 2004. Design, implementation and evaluation of a practical pseudoknot folding algorithm based on thermodynamics. BMC Bioinformatics 5:104.

Rivas E, Eddy SR, 2001. Noncoding RNA gene detection using comparative sequence analysis. BMC Bioinformatics 2:8.

Roberts SB, Emmons SW, Childs G, 1989. Nucleotide sequences of *Caenorhab-*

ditis elegans core histone genes: Genes for different histone classes share common flanking sequence elements. J Mol Biol 206:567–577.

Rodriguez A, Griffiths-Jones S, Ashurst JL, Bradley A, 2004. Identification of mammalian microRNA host genes and transcription units. Genome Research 14:1902–1910.

Schattner P, Brooks AN, Lowe TM, 2005. The `tRNAscan-SE`, `snoscan` and `snoGPS` web servers for the detection of tRNAs and snoRNAs. Nucl Acid Res 33:W686–W689.

Stein DL, Bao Z, Blasia D, Blumenthal T, Brent MR, Chen N, et al., 2004. The genome sequence of *Caenorhabditis briggsae*: A platform for comparative genomics. PLOS Biology 1:167–192.

Storz G, Altuvia S, Wassarman KM, 2005. An abundance of RNA regulators. Annu Rev Biochem 74:199–217.

Stricklin SL, Griffiths-Jones S, Eddy SR, 2005. C. elegans noncoding RNA genes. WormBook doi/10.1895/wormbook.1.7.1. Http://www.wormbook.org/chapters/www_noncodingRNA/noncodingRNA.html.

Szymański M, Barciszewska MZ, Żywicki M, Barciszewski J, 2003. Noncoding RNA transcripts. J Appl Genet 44:1–19.

Thijs G, Lescot M, Marchal K, Rombauts S, De Moor B, Rouz MY, 2001. A higher order background model improves the detection of regulatory elements by Gibbs Sampling. Bioinformatics 17:1113–1122.

Thomas J, Lea K, Zucker-Aprison E, Blumenthal T, 1990. The spliceosomal snRNAs of *Caenorhabditis elegans*. Nucleic Acids Res 18:2633–2642.

Thompson JD, Higgs DG, Gibson TJ, 1994. CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties, and weight matrix choice. Nucl Acids Res 22:4673–4680.

Tycowski KT, Shu MD, Steitz JA, 1996. A mammalian gene with introns instead of exons generating stable RNA products. Nature 6564:464–466.

Tycowski KT, Steitz JA, 2001. Non-coding snoRNA host genes in Drosophila: expression strategies for modification guide snoRNAs. Eur J Cell Biol 80:119–125.

Wagner A, Stadler PF, 1999. Viral RNA and evolved mutational robustness. J Exp Zool MDE 285:119–127.

Wang J, Zhang J, Zheng H, Li J, Liu D, Li H, Samudrala R, Yu J, Wong GK, 2004. Mouse transcriptome: Neutral evolution of 'non-coding' complementary DNAs. Nature 431:1–2.

Washietl S, Hofacker IL, 2004. Consensus folding of aligned sequences as a new measure for the detection of functional RNAs by comparative genomics. J Mol Biol 342:19–30.

Washietl S, Hofacker IL, Stadler PF, 2005a. Fast and reliable prediction of noncoding RNAs. Proc Natl Acad Sci USA 102:2454–2459.

Washietl S, Hofacker IL, Stadler PF, 2005b. Thousands of noncoding RNAs with conserved structure in mammalian genomes. Nature Biotech In press.